DOCUMENT RESUME

ED 260 127 TM 850 457

AUTHOR Schaeffer, Gary A.; And Others
TITLE Assessing the Reliability of Criterion-Referenced Measures Used to Evaluate Health-Education Programs.
SPONS AGENCY Centers for Disease Control (DHHS/PHS), Atlanta, GA.
PUB DATE Oct 84
CONTRACT 200-83-0619
NOTE 21p.; Paper presented at the joint meeting of the Evaluation Network and Evaluation Research Society (San Francisco, CA, October 10-13, 1984). Supported in part by a contract awarded to IOX Assessment Associates, Culver City, CA.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Correlation; *Criterion Referenced Tests; Cutting Scores; Elementary Secondary Education; Error of Measurement; *Health Education; *Mastery Tests; Mathematical Formulas; Norm Referenced Tests; *Program Evaluation; *Test Reliability; Test Theory

ABSTRACT

The reliability of criterion referenced tests, which are often used to evaluate health education programs, may be conceptualized in different ways. Classical conceptualizations of test reliability have limited usefulness when applied to health-related criterion referenced tests. When a cutting score is set, test reliability can be represented as the consistency of mastery versus nonmastery classifications; the Kappa index may also be used. When a cutting score is not set, the size of the standard error of measurement of a domain score estimate reflects test reliability. Five simple mathematical formulas that estimate measurement error for criterion referenced tests are presented. Measurement error may be used to compute confidence intervals for domain score estimates and for the cutting score. (Author/GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED260127

Assessing the Reliability of Criterion-Referenced Measures Used to Evaluate Health-Education Programs

Gary A. Schaeffer
Robert E. Carlson
Robin L. Matas

IOX Assessment Associates
11411 W. Jefferson Blvd.
Culver City, CA 90230

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

B.A Schaeffer

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 850 457

Abstract

The reliability of criterion-referenced tests (CRTs) used in health program evaluations can be conceptualized in different ways. Classical conceptualizations of reliability have limited usefulness when applied to health-related CRTs. When a cut-score is set, the test's reliability can be represented as the consistency of mastery/nonmastery classifications. When a cut-score is not set, the size of the standard error of measurement (SEM) of a domain score estimate is a reflection of the test's reliability. Computationally simple formulas are presented that estimate appropriate SEMs for CRTs. The SEM can be used in computing confidence intervals for domain score estimates and for the cut-score if there is one.

Assessing the Reliability of Criterion-Referenced Measures Used to Evaluate Health-Education Programs[1]

Accountability is a catchword that is being applied in most professions, including health education. As such, it is becoming imperative that health-education professionals document the successful and less successful components of their programs. With program resource dollars shrinking, program accountability data become prominent in the decision to continue or discontinue a program. Also, accountability data may be used to isolate specific weaknesses in a program with the goal of improving the program in the future. It is clear, then, that accountability of programs must be done accurately.

One major way of documenting the successes and failures of health-education programs is by administering measures to program participants during various phases of the program. (Measures may be tests, questionnaires, surveys, observations, and other assessment instruments. The words "test" and "measure" will be used interchangeably in this report.) These measures may assess health-related skills, knowledge, attitudes, or behaviors. The measures are designed to reflect the status of participants with respect to these health-related goals. If, for example, participants' knowledge increases over the course of the program,

then data from a knowledge measure can serve as an indicator of a program's success. The accuracy of these data typically is reported in terms of reliability.

Measures used in health-program evaluations often are criterion-referenced rather than norm-referenced. Norm-referenced measures traditionally have been used in assessing intellectual capabilities. The major purpose of norm-referenced measures is to obtain a rank order of examinees; that is, to accurately determine how examinees perform compared with each other. Traditional reliability coefficients assess the consistency of the rank order of examinees over similar testing occasions.

Criterion-referenced tests (CRTs), on the other hand, are designed to assess examinee performance on a domain of tasks irrespective of how other examinees perform. CRTs may be used to monitor individual progress in objectives-based instructional programs. For example, a CRT of knowledge of the effects of cigarette smoking would be used to determine the extent to which students have acquired that knowledge. Comparisons of students' performance with the performances of other students are not important. What is important is the absolute magnitude of students' scores, and the consistency of these scores over multiple test administrations. Methods of assessing CRT reliability relevant to this assessment purpose are available. The methods are different from the traditional ones used to assess reliability of norm-referenced tests.

There are four sections in this report. First will be a discussion of how norm-referenced approaches to test reliability

are insufficient or inappropriate for assessing the reliability of CRTs. The next section will present indices useful in assessing reliability of a CRT when a cut-score is set to differentiate "masters" from "nonmasters" of the domain of interest. The third part will present the standard error of measurement as a useful index of a measure's reliability when no cut-score is set. Finally, summary statements will be presented.

Problems with Traditional Reliability Indices

Traditional approaches to reliability, namely test-retest, parallel forms, and internal consistency, involve correlations. Internal-consistency reliability can be thought of as the average correlation between all possible split-half forms, when items on the measure are divided into two half-length measures. Test-retest and parallel-forms reliability traditionally are reported as correlations of test scores between two test administrations for a group of individuals.

Several problems arise when these correlational indices of reliability are applied to CRTs. The test-retest and parallel-forms correlations assess the extent to which examinees on the first test administration are ranked similarly on the second test administration. The consistency of rank order, however, does not necessarily reflect the reliability of a CRT.

To take an exaggerated case to illustrate the point, suppose three examinees scored, respectively, 82, 83, and 84 correct out of 100 items on the first test administration, and then scored, respectively, 93, 94 and 95 correct on the second test administration. The test-retest or parallel-forms reliability coefficient

would be the correlation of the two sets of scores, or 1.0. The test would be said to be highly reliable in the traditional sense of reliability. Notice, however, that the actual values of the scores changed; that is, each examinee improved by eleven items. This discrepancy in scores is not reflected in the reliability correlation. With CRTs, however, the absolute magnitude of the score differences over testing occasions is of primary interest.

Also, the reverse may be true, that is, a measure may be unreliable in the traditional sense but be perfectly reliable in the CRT sense. For example, suppose that on the first test administration three examinees score 70, 71, and 72, respectively, and on the second test administration they score 71, 72, and 70, respectively. This test would not be reliable in the traditional sense -- the correlation of scores between the two test administrations would be very low. This measure would be considered reliable in the CRT sense, however, because the absolute magnitudes of the differences in scores from the two test administrations are small. Of course, an actual reliability study would consist of many more than three individuals.

A second problem with applying correlational techniques to CRTs is related to test score variance. In order for correlations to be meaningful, there must be a sufficient amount of test score variance. A set of examinee scores with little or no variance will necessarily lead to correlation coefficients at or near zero. In the traditional sense this is accurate because the test is not able to spread out, or rank order, examinees.

CRTs, however, often yield limited test score variance by design, and the traditional reliability interpretation would be

misleading. For example, a CRT may be designed to assess each examinee with respect to some domain of health-related knowledge. This assessment is not to be made relative to the performance of other examinees. Rather, this assessment is based on the actual score obtained by each examinee. After a successful program concludes, most of the participants will have acquired the knowledge taught in the program. A test of this knowledge at the end of the program will yield limited test score variance, and hence low reliability in the traditional sense. This test, however, may be quite reliable in the CRT sense.

Internal-consistency reliability presents additional concerns. In fact, the appropriateness of internal-consistency reliability estimates in assessing consistency of measurement in the CRT framework has been questioned (Popham, 1978). Internal consistency, probably the most frequently reported reliability statistic for norm-referenced tests, does not deal with the reliability of total test scores. Rather, it deals only with characteristics of the items themselves. Internal consistency, therefore, might better be considered as a means of verifying the homogeneity of CRT items rather than as a measure of CRT reliability.

Traditional reliability methods, therefore, have problems when applied to CRTs. Assessing the reliability of CRTs requires different conceptualizations of reliability. The following sections will present methods of assessing CRT reliability.

Reliability When A Cut-Score Is Set

Sometimes in health program evaluation a test score is set such that examinees who score above it are considered "masters" and examinees who score below it are considered "nonmasters." Mastery status, of course, only refers to the domain being measured. For example, a curriculum designed to affect children's attitudes toward smoking cigarettes may be assessed by an attitudinal questionnaire. The instructor may set a cut-score such that children who score at or above the cut-score are considered to be masters (i.e., have "appropriate" attitudes) and those who score below the cut-score are considered to be nonmasters (i.e., have "inappropriate" attitudes). Perhaps children with inappropriate attitudes will be given additional curricular activities.

When a cut-score is set to distinguish masters from nonmasters, the primary reliability question is the following: How consistently are individuals classified as masters or nonmasters on repeated measurements of one test form or parallel forms?[2] That is, are examinees who are classified as masters (or nonmasters) on one test administration consistently classified as masters (nonmasters) on another test administration? In health-program evaluation it is frequently necessary to have at least two parallel forms of a test. Often, one form is administered before the program begins, and the other form is administered at the end of the program. Several reliability

indices have been proposed for assessing classification consistency (see Berk, 1980; and Hambleton, Swaminathan, Algina, and Coulson, 1978 for lists of these indices). Most of the indices assume that a dichotomous classification based on a cut-score has been made, and that the errors associated with all false mastery and false nonmastery classifications are equally serious (i.e., it is just as serious an error to classify a true master as a nonmaster as it is to classify a true nonmaster as a master.)

One index is the proportion of individuals consistently classified as masters or nonmasters on repeated measurements using one form or parallel forms (Hambleton and Novick, 1973). This index incorporates the measurement accuracy of the test itself as well as the mastery/nonmastery composition of the group tested. The index, however, is sensitive to the position of the cut-score. Low values of this index tend to occur with cut-scores near the mean and high values usually are associated with cut-scores near a tail of most score distributions. The index also is affected by classifications that occur by chance. That is, even with a completely inaccurate test, some classifications would be consistent over two test administrations.

Another index, called kappa (Cohen, 1960), deletes the proportion of consistent decisions expected by chance from the proportion actually observed. Kappa can be interpreted as the proportion of consistent classifications beyond that expected by chance. The interpretation of kappa is complicated, however, because the range of kappa may be restricted by test length, score variability, and marginal frequencies. In fact, kappa can only equal 1.0 when both forms or measurements report the same

numbers of masters (and nonmasters).

In practice, both the classification consistency index and kappa should be computed to reflect the reliability of a CRT when a cut-score is set. They yield different information and are simple to compute. They should, however, be interpreted with the caveats listed above. Formulas for these indices can be found in the Appendix.

Reliability When A Cut-Score Is Not Set

Although cut scores are occasionally used in health program evaluations, more often cut-scores are not set. In such instances, attention is focused in estimating how much of the domain of interest program participants have mastered. The domain consists of all possible items that measure what a test is designed to measure. Each test form may be thought of as a random sample of items from the domain. For example, a health-education curriculum component for elementary school students may deal with proper decision-making techniques. Questionnaire items consisting of vignettes may ask students to select from alternatives the next step that should be taken in making a decision. These vignettes may be thought of as a random sample of all possible vignettes that assess students' decision-making skills. The instructor is not just interested in how students would respond to the particular vignettes that appear on the test. Rather, interest is in generalizing performance on the test to the domain of vignettes.

Some may object to the assumption of randomly parallel tests because tests are not usually built by a strict process of random sampling from a domain of items. Even so, the assumption of randomly

parallel tests is a useful idealization. Although the items of a particular test may not have been drawn at random, one can usually conceive of a domain of items from which the test items might have been drawn.

The primary reliability question in instances where no cut-score is set is the following: How accurately can a test estimate examinees' domain scores? An estimate of the domain score, when each item is scored either correct or incorrect, is the proportion of correct answers obtained, or the p-value. For example, if an examinee, correctly answers 16 items on a 20-item test, then that individual's score is 16/20 for a p-value of .80. This p-value represents an unbiased estimate of the examinee's performance on the entire domain of items. Similarly, on a measure where items are scored on, for example, a 5-point scale, the domain score estimate is the score obtained divided by the highest score possible.

To compute the reliability of a test that may be considered a random sample of items from a domain, the principles of generalizability theory may be used (Cronbach, Gleser, Nanda, and Rajaratnam, 1972). Reliability estimates based on generalizability theory usually will be different from those based on classical test theory from which traditional reliability indices are derived. Differences are found in the derivations of the error variances. In classical test theory, the error variance is used to make comparative or relative decisions about persons. By contrast, the error that is appropriate for randomly parallel CRTs deals with decisions about the absolute magnitude of persons' scores (Brennan, 1980).

The error variance from generalizability theory, $\sigma^2(\Delta)$, is the average within-person error variance, where error is defined as the difference between a person's observed and domain scores. Generalizability theory recognizes multiple sources of error. The influence of various sources of error is estimated by applying analysis of variance methodology to data collected under various designs. It is especially important to note that includes the sampling variability of the main effect for sets of items, or tests, whereas the classical definition of error variance does not. Consequently, $\sigma^2(\Delta)$ is always at least as large as the classical error variance. They are equal only when the mean scores for two randomly parallel tests are the same. This condition holds when tests are classically parallel.

There is a generalizability coefficient that is analogous to a classical reliability coefficient. It is interpretable in much the same way as the classical estimate of internal consistency, even though the derivation is based on a different model. Another related generalizability coefficient is computed with $\sigma^2(\Delta)$ rather than the classical error variance component. These generalizability coefficients, however, do not indicate the relative magnitude of different sources of error, nor do they indicate how to improve the measurement procedure. In fact, the specific application of these generalizability coefficients in practice is unclear (Berk, 1980).

Values of the error variances derived from generalizability theory, however, are useful in establishing confidence intervals about the domain score estimates. The size of these confidence intervals is a reflection of the test's reliability. More reli-

able tests result in smaller confidence intervals. In classical test theory the standard error of measurement (SEM), that is, the square root of the error variance, may be used to establish confidence intervals about an individual's observed score. It may be said that, for example, 95% of the individuals with a certain true score would be expected to score between ±2 standard errors of measurement about their true score. This principle also applies to CRTs, but the appropriate error variance is derived differently.

There are computationally simple formulas that may be used to estimate the standard error of measurement both for an individual and for a population of individuals when each individual is administered only one form of a CRT. For example, if p is the proportion correct on a test consisting of n items, then the standard error of measurement (i.e., the square root of the error variance) for individual i is (Lord, 1957):

$$(1)\quad SEM(i) = \sqrt{\frac{p(1-p)}{n-1}}$$

Although the p-values are drawn from a binomial distribution, approximate confidence intervals may be obtained by the use of normal deviates. For example, the SEM for individuals with p = .80 when n = 20 would be .09. Thus, the 95% confidence interval would be .62 - .98.

Note that the size of the confidence interval may be different for other persons because the value of the standard error of measurement depends on p.

If test scores are reported in raw score form (r), then equation (1) becomes:

$$(2)\quad SEM(x) = \sqrt{\frac{r(n-r)}{n-1}}$$

A single standard error of measurement for a population of examinees is the square root of the average error variance for each examinee. This is equivalent to the square root of . This standard error of measurement can be estimated with one test form administration using the following formula (Millman, 1979):

$$(3)\quad SEM(pop) = \sqrt{\frac{\bar{X}(1-\bar{X}) - S^2(X)}{n-1}}$$

where $\bar{X}$ = mean of scores on the test expressed as proportion correct

n = number of items

$S^2(X)$ = sample variance of test scores expressed as proportion correct

Simple formulas also exist for estimating the standard error of measurement in situations where individuals have been administered the test twice (either test-retest or parallel-forms). To calculate the standard error of measurement for an individual who has taken two test forms, the following formula can be used (Lord and Novick, 1968):

$$(4)\quad SEM(x) = \frac{|d(x)|}{\sqrt{2}}$$

where $|d(i)|$ = the absolute value of the difference in observed total scores from one form to another for examinee i.

It should be noted that for a particular test form, the errors of measurement are usually biased, that is, the errors tend to be consistently negative or positive depending on whether the particular test form is more difficult or less difficult than the other test forms.

For the population of N individuals, the standard error of measurement when two parallel test forms have been administered to each examinee is (Lord and Novick, 1968):

$$(5)\quad SEM(pop) = \sqrt{\frac{1}{2N}\sum_{i=1}^{N} d^2(i)}$$

This also is an estimate of the square root of $\sigma^2(\Delta)$.

The standard error of measurement for an individual score and the standard error of measurement for a population of individuals have different applications (Berk, 1980). Each reflects a different aspect of the reliability of a CRT. The SEM for an individual score can be placed around a cut-score when one is set. Thus, if a cut-score is set at 70%, then p = .70 could be used in equation (1). If a 95% confidence interval is put around a cut-score, then the probabilty of correct mastery and non-mastery decisions will be at least 95%. For other uses, such as placing confidence intervals about individuals' test scores or in describing the accuracy of the test as a whole, Berk (1980) recommends that estimates of $\sigma^2(\Delta)$, that is, estimates for the population of examinees, be used. These estimates can be calculated using equations (3) or (5).

Summary Statements

Several points have been presented in this report. They are listed below as summary statements.

1. Measures used in health-program evaluations typically are criterion-referenced rather than norm-referenced.
2. Classical conceptualizations of reliability have limited usefulness when applied to CRTs.
3. When a cut-score is set, reliability of a test can be represented as the consistency of mastery/nonmastery classifications.
4. Consistency of mastery/nonmastery classifications, however, is affected by chance. An index such as kappa that takes chance into account should be reported with the classification consistency index.
5. When a cut-score is not set, the absolute magnitude of the domain score estimate should be the focus of attention.
6. The size of the standard error of measurement of a domain score estimate is a reflection of test reliability. Assuming randomly parallel tests, the square root of the error variance derived in generalizability theory is an appropriate standard error of measurement.
7. Computationally simple formulas may be used to estimate standard errors of measurement in instances where examinees take one or two randomly parallel forms of a test.
8. The standard error of measurement can be used in computing confidence intervals about domain scores and about the cut-score if there is one.

References

Berk, R.A. (1980). A consumer's guide to criterion-referenced test reliability. Journal of Educational Measurement, 17, 323-349.

Brennan, R.L. (1980). Applications of generalizability theory. In R.A. Berk (Ed.) Criterion-referenced measurement: The state of the art. Baltimore: Johns Hopkins University Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Hambleton, R.K., & Novick, M.R. (1973). Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 10, 159-170.

Hambleton, R.K., Swaminathan, H., Algina, J., & Coulson, D.B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 48, 1-47.

Lord, F.M. (1957). Do tests of the same length have the same standard errors of measurement? Educational and Psychological Measurement, 17, 510-521.

Lord, F.M., & Novick, M.R. (1968). Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley.

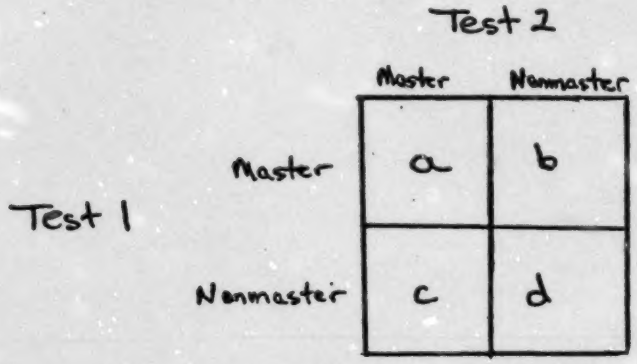
Millman, J. (1979). Reliability and validity of criterion-referenced test scores. In R.E. Traub (Ed.), New Directions for Testing and Measurement. San Francisco: Jossey-Bass.

Popham, W.J. (1978). Criterion-referenced measurement. Englewood Cliffs, N.J.: Prentice-Hall.

Subkoviak, M.J. (1980). Decision-consistency approaches. In R.A. Berk (Ed.) Criterion-referenced measurement: The state of the art. Baltimore: Johns Hopkins University Press.

Appendix

Listed below are procedures for computing classification consistency. Each examinee is tested twice, and a cut-score is set to differentiate masters and nonmasters.



$$P_o = \frac{a+d}{N}$$

$$P_c = \frac{(a+b)(a+c) + (c+d)(b+d)}{N^2}$$

$$\text{Kappa} = \frac{P_o - P_c}{1 - P_c}$$

where: a, b, c, and d are the numbers of examinees in each cell

N = the total number of examinees

P_o = the observed classification consistency index

P_c = the proportion of agreement expected by chance

Kappa = the proportion of agreement above that expected by chance

Footnotes

1
This research was supported in part by a contract awarded by the Centers for Disease Control, Atlanta, Georgia, to IOX Assessment Associates, Culver City, California (Contract #200-83-0619). The responsibility for the contents of the article, however, belongs solely to the authors themselves.

2
There are methods for estimating classification consistency based on only one test administration. These methods, however, are computationally tedious and provide biased estimates for short tests (Subkoviak, 1980).